

# New Discriminative Methods for RNA Gene-finding and Classification

Richard F. Meraz<sup>1</sup>, Xiaofeng He<sup>2</sup>, Chris H.Q. Ding<sup>2</sup>, Yan Karklin<sup>1</sup>, Stephen R. Holbrook<sup>1</sup>

<sup>1</sup>Physical Biosciences Division and <sup>2</sup>Computer Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA

## Summary

RNA genes lack most of the signals used for protein gene identification. The major shortcoming of previous discriminative methods to distinguish functional RNA (tRNA) genes from other non-coding genomic sequences is that only positive examples of tRNAs are known; there are no confirmed negatives -- only intergenic sequences that may be positive or negative. To address this problem we developed the "Positive Sample Only Learning" (PSOL) method. Unlabeled sequences are searched in parameter space for a group that is far away from the positive set and mutually far away among themselves. Large Margin Classifiers are iteratively trained to enlarge this negative set. The remaining examples are putative tRNA genes.

To group uncharacterized tRNA genes into meaningful structural and functional classes we have extended a dual graph representation<sup>4</sup> of RNA secondary structure to include labels for loops, helices, and the lengths of these features. A kernel function is defined directly on these structures for automatic class discovery based on topological similarities in secondary structure.

These methods are evaluated by revisiting the problem of RNA gene-finding in the *E. coli*/K12 genome.

Their performance is evaluated by revisiting the problem of RNA gene-finding in the *E. coli*/K12 genome.

## Positive Sample Only Learning (PSOL)

### Initial selection of negative set

The point set distance  $d(x_i, P)$  is defined as the minimum distance between a sequence in parameter space  $x_i$  and the positive set of parameterized known tRNA sequences  $P$ . Given several points in the current negative set  $N$ : a new point  $x_j$  is selected based on maximum dissimilarity to the positive set and the maximum distance to the negative set -- (corresponding to d' and s' in the figure):

$$\max_{x_j \in (x_i, N)} \left[ d(x_j, P) \sum_{x_k \in N} d(x_i, x_k) \right]$$

Once a specified size of  $N$  is reached, the algorithm terminates and we set the initial negative training set  $N_{train} = N$ .

### Negative set expansion

We train a Support Vector Machine (SVM) on the dataset  $P + N_{train}$  to obtain a large margin decision boundary. Denote the support vectors in  $N_{train}$  for this SVM as  $\lambda_{N_{train}}$ . Each point in the unlabeled parameterized intergenic sequences  $U$  will have an SVM decision value  $f(x_i)$ . We apply a threshold  $h$  to select points away from the boundary:

$$N_h = \{x_i \mid f(x_i) \leq -h, h > 0\}$$

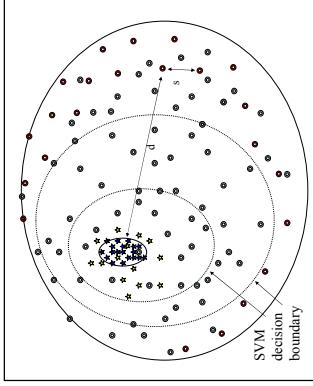
To balance the positive and negative sets, we limit the size of newly predicted negative samples  $N_{pred}$  to be  $r|P|$ . Assume the decision values  $f(x_i)$  are sorted in increasing order, then

$$N_{pred} = \{x_i \text{ s.t. } i \leq r|P| \text{ and } f(x_i) \leq -h\}$$

The total negative set becomes  $N_{neg} = N_{pred} \cup N_{pred}$ , unlabeled set  $U = U - N_{pred}$ , and the current negative training set  $N_{train} = N_{pred} + N_{pred}$ .

Negative expansion is repeated until either (a) convergence or (b) the size of the current unlabeled is equal to a predetermined number of positive samples.

## Analyzing Predictions of tRNA genes in *E. coli* K12 from PSOL and Previous Large-Scale Screens



- (1) Fifty-six tRNA and a compliment of tRNA and tRNA sequences were split into 499 overlapping windows of a length of 80 and overlap of 40 nucleotides.
- (2) Intergenic regions were split the same way into 11542 windows to comprise the unlabeled set.
- (3) Sequences parameterized as sequence-statistics and similarity scores to bacterial genomes were inputs to the PSOL method.
- (4) PSOL averaged **85.51 ± 3.5%** recovery of hold-out positive examples over 5 repeats of a 5-fold hold-out experiment.
- (5) The histogram above shows the final SVM after convergence of the PSOL method. The dashed line at 0 is the decision boundary.

## Kernels on Labeled Dual Graphs for RNA Classification

